

NORMS OF RANDOM MATRICES: LOCAL AND GLOBAL PROBLEMS

ELIZAVETA REBROVA AND ROMAN VERSHYNIN

ABSTRACT. Can the behavior of a random matrix be improved by modifying a small fraction of its entries? Consider a random matrix A with i.i.d. entries. We show that the operator norm of A can be reduced to the optimal order $O(\sqrt{n})$ by zeroing out a small submatrix of A if and only if the entries have zero mean and finite variance. Moreover, we obtain an almost optimal dependence between the size of the removed submatrix and the resulting operator norm. Our approach utilizes the cut norm and Grothendieck-Pietsch factorization for matrices, and it combines the methods developed recently by C. Le and R. Vershynin and by E. Rebrova and K. Tikhomirov.

1. INTRODUCTION

1.1. Local and global problems. When a certain mathematical or scientific structure fails to meet reasonable expectations, one often wonders: is this a local or global problem? In other words, is the failure caused by some small, localized part of the structure, and if so, can this part be identified and repaired? Or, alternatively, is the structure entirely, globally bad? Many results in mathematics can be understood as either local or global statements. For example, not every measurable function $f : \mathbb{R} \rightarrow \mathbb{R}$ is continuous, but Lusin's theorem implies that f can always be made continuous by changing its values on a set of arbitrarily small measure. Thus, imposing continuity is a local problem. On the other hand, a continuous function may not be differentiable, and there even exist continuous and nowhere differentiable functions. Thus imposing differentiability may be a global problem. In statistics, the notion of *outliers* – small, pathological subsets of data, the removal of which makes data better – points to local problems.

1.2. Random matrices and their norms. In this paper about random matrices we ask: is bounding the norm of a random matrix a local or a global problem? To be specific, we consider $n \times n$ random matrices A with independent and identically distributed (i.i.d.) entries. The *operator norm* of A is defined by considering A as a linear operator on \mathbb{R}^n equipped with

R. V. is partially supported by NSF grant 1265782 and U.S. Air Force grant FA9550-14-1-0009.

the Euclidean norm $\|\cdot\|_2$, i.e.

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

Suppose for a moment that the entries of A have zero mean and bounded fourth moment, i.e. $\mathbb{E} A_{ij}^4 \leq C$ where C is a constant. Then

$$\|A\| = O(\sqrt{n}) \tag{1.1}$$

with high probability. This is the consequence of Bai-Yin's law, which is valid as $n \rightarrow \infty$ [2]. Non-asymptotic versions of this bound, which hold for fixed dimensions n , can be deduced from [16, 8, 5]. Note that the $O(\sqrt{n})$ bound is the best one can generally hope for. Indeed, if the entries of A have unit variance, then the typical magnitude of the Euclidean norm of a row of A is $\sim \sqrt{n}$, and the operator norm of A can not be smaller than that. Moreover, the entries of A must have finite fourth moment for (1.1) to hold [4].

1.3. Main results. Now let us postulate nothing at all about the distribution of the i.i.d. entries of A . It still makes sense to ask: *is enforcing the ideal bound (1.1) for random matrices a local or a global problem?* That is, can we enforce the bound (1.1) by modifying the entries in a small submatrix of A ? We will show in this paper that this is possible if and only if the entries of A have zero moment and finite variance. The “if” part is covered by the following theorem.

Theorem 1.1 (Local problem). *Consider an $n \times n$ random matrix A with i.i.d. entries that have mean zero and unit variance, and let $\varepsilon \in (0, 1/2]$. Then, with probability at least $1 - 7\exp(-\varepsilon n/4)$, there exists an $\varepsilon n \times \varepsilon n$ submatrix of A such that replacing all of its entries with zero leads to a well-bounded matrix \tilde{A} :*

$$\|\tilde{A}\| \leq \frac{C \ln \varepsilon^{-1}}{\sqrt{\varepsilon}} \cdot \sqrt{n},$$

where C is a sufficiently large absolute constant.

Remark 1.2 (Optimality). The dependence on ε in Theorem 1.1 is best possible up to the $\ln \varepsilon^{-1}$ factor. To see this, let $p := 2\varepsilon/n$ and suppose A_{ij} take values $\pm 1/\sqrt{p}$ with probability $p/2$ each and value 0 with probability $1-p$. Then A_{ij} have zero mean and unit variance as required. The expected number of non-zero entries in A equals $pn^2 = 2\varepsilon n$. Thus the number of the rows of A containing these entries is bigger than εn with high probability. (This is a standard observation about the balls-into-bins model.) Therefore, no $\varepsilon n \times \varepsilon n$ submatrix can contain all the non-zero entries of A . In other words, \tilde{A} must contain at least one non-zero entry of A , and thus it has magnitude

$$\|\tilde{A}\| \geq \frac{1}{\sqrt{p}} \gtrsim \frac{\sqrt{n}}{\sqrt{\varepsilon}}.$$

This shows that the dependence on ε in Theorem 1.1 is almost optimal.

By rescaling, a more general version of Theorem 1.1 holds for any finite variance of the entries. The two main assumptions in this theorem – mean zero and finite variance – are necessary in Theorem 1.1. Without either of them, the problem becomes global in a strong sense: the desired $O(\sqrt{n})$ bound can not be achieved even after modifying of a *large* submatrix. This is the content of the following result.

Theorem 1.3 (Global problem). *Consider an $n \times n$ random matrix A_n whose entries are i.i.d. copies of a random variable that has either nonzero mean or infinite second moment,¹ and let $\varepsilon \in (0, 1)$. Then*

$$\min \frac{\|\tilde{A}_n\|}{\sqrt{n}} \rightarrow \infty \quad \text{as } n \rightarrow \infty$$

almost surely. Here the minimum is with respect to the matrices \tilde{A}_n obtained by any modification of any $\varepsilon n \times \varepsilon n$ submatrix of A_n .

It should be noted that while Theorem 1.1 becomes harder for smaller ε , Theorem 1.3 becomes harder for larger ε , those near 1.

We prove Theorem 1.3 in Section 9. The argument is considerably simpler than for Theorem 1.1. Indeed, the nonzero mean forces the sum of the entries of \tilde{A}_n to be $\gtrsim n^2$, and the infinite second moment forces the Frobenius norm of \tilde{A}_n (the square root of the sum of the entries squared) to be $\gg n^2$ with high probability. Either of these two bounds can be easily used to show that the operator norm of \tilde{A}_n is $\gg \sqrt{n}$.

1.4. Related results. There have been several precursors to this work. When Y. Yin, Z. Bai, P. Krishnaiah and J. Silverstein showed that $\|A\| \gg \sqrt{n}$ if the i.i.d. entries of A have infinite weak fourth moment [4], they obtained this result by checking that the largest entry of A must be $\gg \sqrt{n}$ in this case. However, the number of such large entries is typically small. This suggests – but does not prove – that the only obstruction to the desired bound $\|A\| = O(\sqrt{n})$ could be a few large entries, and that removing those entries could enforce this bound. This is consistent with the conclusion Theorem 1.1.

A related result for the partial case of a symmetric Bernoulli random matrix B was proved by U. Feige and E. Ofek [6]; an alternative argument and a more informative result was given later in [10]. Suppose the entries of B on and above the diagonal are independent, Bernoulli random variables with mean $p \in (0, 1)$. If one removes the heavy rows and columns – those containing more than $2pn$ ones, then the resulting matrix B' satisfies the optimal norm bound $\|B' - \mathbb{E} B'\| \leq C\sqrt{pn}$. (To see that this bound is

¹Although this is a minor terminological distinction, in this theorem we prefer to talk about second moment rather than variance. This is because the second moment $\mathbb{E} X^2$ of a random variable X is always defined in the extended real line, while the variance $\text{Var}(X) = \mathbb{E}(X - \mathbb{E} X)^2$ is undefined if the mean $\mathbb{E} X$ is infinite.

consistent with that of Theorem 1.1, divide both sides by \sqrt{p} to normalize the variance of the entries.) One can quickly check using concentration that the number of heavy rows and columns in B is typically small. With a little more work, one can even place all ones from the heavy rows and columns into a small submatrix (see Lemma 8.1 below). Thus Feige-Ofek's result is an example of Theorem 1.1.

Weaker versions of Theorem 1.1, with an additional factor $\log n$ in the norm bound and weaker probability guarantees, can be derived from known general bounds on random matrices, such as the matrix Bernstein's inequality [17]. (One would apply the matrix Bernstein's inequality for the entries truncated at level \sqrt{n} , and control the larger entries as in Section 8.) A different weaker bound $\|\tilde{A}\| \leq (C/\varepsilon)\sqrt{n}$, which has a suboptimal dependence on ε , can be derived in a faster way by using results of [13] directly rather by improving the method of [13].

2. THE METHOD

Our approach to Theorem 1.1 utilizes and advances the methods developed recently in [13] and [10]. We will first control the cut norm of A and then pass to the operator norm using Grothendieck-Pietsch factorization. Let us describe these steps in more detail.

2.1. Three matrix norms. The operator norm of a matrix A , as we already mentioned, is defined by considering A as a linear operator on the (finite dimensional) space ℓ_2 , i.e.

$$\|A\| = \|A : \ell_2 \rightarrow \ell_2\|.$$

Rather than bounding the operator norm of a random matrix A directly, we shall compare it with two simpler norms,

$$\|A\|_{\infty \rightarrow 2} = \|A : \ell_\infty \rightarrow \ell_2\| = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_\infty}$$

and

$$\|A\|_{2 \rightarrow \infty} = \|A : \ell_2 \rightarrow \ell_\infty\| = \max_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_2}.$$

The simplest of the three is the $2 \rightarrow \infty$ norm. A quick check reveals that it equals the maximum Euclidean norm of the rows A_i^T of A :

$$\|A\|_{2 \rightarrow \infty} = \max_{i \in [n]} \|A_i\|_2. \quad (2.1)$$

The next simplest norm is $\infty \rightarrow 2$, which can be conveniently computed as

$$\|A\|_{\infty \rightarrow 2} = \max_{x \in \{-1, 1\}^n} \|Ax\|_2. \quad (2.2)$$

This norm is equivalent within a constant factor to the *cut norm* from the computer science literature [3, 1], where the maximum is taken over $\{0, 1\}^n$.

The hardest of the three is the operator norm,

$$\|A\| = \max_{x \in S^{n-1}} \|Ax\|_2. \quad (2.3)$$

To see why the difficulty in bounding these norms rises this way, note that one has to control n random variables in (2.1), 2^n random variables in (2.2), and infinitely many random variables in (2.3).

2.2. Ideal relationships among the norms. How large do we expect the three norms to be for random matrices? For a simple example, let us first consider a Gaussian random matrix A with i.i.d. $N(0, 1)$ entries. Then it is not difficult to check that

$$\|A\|_{2 \rightarrow \infty} \sim \sqrt{n}, \quad \|A\|_{\infty \rightarrow 2} \sim n, \quad \|A\| \sim \sqrt{n}. \quad (2.4)$$

Indeed, note that the rows of A have Euclidean norms \sqrt{n} on average, so the bound on the $2 \rightarrow \infty$ norm follows by union bound and using Gaussian concentration. The bound on the $\infty \rightarrow 2$ norm follows from (2.2) by using Gaussian concentration for the normal random vector Ax and taking the union bound over $\{-1, 1\}^n$. The bound on the operator norm is a non-asymptotic version of Bai-Yin's law, see e.g. [18, Theorem 5.32].

One might wonder if (2.4) holds not only in the Gaussian case but generally for random matrices A with i.i.d. entries that have zero mean and unit variance. In particular, it would be wonderful if the three norms were always related to each other as follows:

$$\|A\| \lesssim \frac{\|A\|_{\infty \rightarrow 2}}{\sqrt{n}} \lesssim \|A\|_{2 \rightarrow \infty} \lesssim \sqrt{n}. \quad (2.5)$$

This, however, would be too optimistic to expect, since the bound $\|A\| \lesssim \sqrt{n}$ can not hold without higher moments assumptions as we mentioned in Section 1.1. Nevertheless, we will obtain a version of (2.5) after removal a small fraction of rows of A . With high probability, we will be able to find subsets of rows $J_1 \subset J_2 \subset J_3$ with cardinalities $|J_i| \leq \varepsilon n$ and such that

$$\|A_{J_3^c}\| \lesssim \frac{\|A_{J_2^c}\|_{\infty \rightarrow 2}}{\sqrt{n}} \lesssim \|A_{J_1^c}\|_{2 \rightarrow \infty} \lesssim \sqrt{n}. \quad (2.6)$$

where the inequalities hide a factor that depends on ε .

2.3. A roadmap of the proof. The first step in proving (2.6) is to find a small set J_1 with $|J_1| \lesssim \varepsilon n$ and such that

$$\|A_{J_1^c}\|_{2 \rightarrow \infty} \lesssim \sqrt{n} \quad (2.7)$$

with high probability. In other words, we would like to bound all rows of A simultaneously by $O(\sqrt{n})$ after removing a few columns of A . To show this we first focus on one row, where we need to bound a sum of independent random variables (the squares of the row's entries). In Theorem 4.2 we show how to bound sums of independent random variables almost surely by gently *damping* the summands. Damping, or reweighting down, is a softer operation than removing entries. It allows us to treat in Section 5

all columns simultaneously without much effort, thus proving (2.7). The argument in this step is similar to the approach proposed recently in [13]. We somewhat simplify the method of [13] and also improve the dependence between the number of removed columns and the resulting $2 \rightarrow \infty$ norm; this will ultimately lead to the optimal dependence on ε in Theorem 1.1.

At the next step, we extend J_1 to a bigger set of rows J_2 with $|J_2| \lesssim \varepsilon n$ and so that

$$\|A_{J_2^c}\|_{\infty \rightarrow 2} \lesssim n. \quad (2.8)$$

Suppose for a moment that we are not concerned about removal of any columns. It is not too hard to show the general bound

$$\mathbb{E} \|A\|_{\infty \rightarrow 2} \lesssim \sqrt{n} \mathbb{E} \|A\|_{2 \rightarrow \infty}, \quad (2.9)$$

for a random matrix A with independent, mean zero entries; we prove this in Lemma 6.1. However, this bound is not very helpful in our situation. We need to work with the matrix $A_{J_1^c}$ instead of A , which is not trivial: the removal of the columns in J_1 that we did in the first step made the entries of $A_{J_1^c}$ dependent. In Lemma 6.2, we first prove a variant of (2.9) for $A_{J_1^c}$ under an additional symmetry assumption on the distribution of the entries of A . Then we manage to remove this assumption with a delicate symmetrization argument, which we develop in the rest of Section 6, with the final result being Theorem 6.6. The general idea of this step, as well as some of our arguments here, are inspired by [13]. However we need to be considerably more careful than in [13] to obtain (2.8) with a *logarithmic* dependence on ε .

Next, we pass from $\infty \rightarrow 2$ norm to the operator norm in Section 7. This is done by using Grothendieck-Pietsch factorization (Theorem 7.1), a result that yields the first inequality in (2.6) for completely arbitrary, even non-random, matrices. This reasoning was recently used in a similar context in [10].

The argument we just described works under the additional assumption that the entries of A be $O(\sqrt{n})$ almost surely. To be specific, such boundedness assumption is needed to make the damping argument in Step 1 work with mild, logarithmic dependence on ε . The contribution of the entries that are larger than \sqrt{n} are controlled in Section 8 by showing that there can not be too many of them. The unit variance assumption implies that there are $O(1)$ such large entries per column on average. This does not mean, of course, that all columns will have $O(1)$ large entries with high probability; in fact there could be columns with $\sim \log n / \log \log n$ large entries. But we will check in Lemma 8.1 that the number of such heavy columns is small; removing them will lead to the desired bound $O(\sqrt{n})$ on the operator norm for the matrix with large entries. We develop this argument in Proposition 8.4 and Corollary 8.6, and derive the full strength of Theorem 1.1 in Section 8.4.

Theorem 1.3 is proved in Section 9. The paper is concluded with Section 10 where we discuss some further problems.

Acknowledgements. We are thankful to Ramon van Handel who showed us a simple argument that we use here to prove Lemma 3.1.

3. PRELIMINARIES

3.1. Notation. Throughout the paper, positive absolute constant are denoted C, C_1, c, c_1 , etc. Their values may be different from line to line. We often write $a \lesssim b$ to indicate that $a \leq Cb$ for some absolute constant C .

The discrete interval $\{1, 2, \dots, n\}$ is denoted by $[n]$. If \mathcal{R} is some subset of indices, $\mathcal{R} \subset [n] \times [n]$, let us denote by $A_{\mathcal{R}}$ the matrix obtained from A by replacing the indices in \mathcal{R} by zero:

$$A_{\mathcal{R}} := (\bar{A}_{ij})_{i,j=1}^n, \text{ where } \bar{A}_{ij} = A_{ij} \mathbb{1}_{\{(i,j) \in \mathcal{R}\}}.$$

We will often consider subsets of columns of the matrix, so when $\mathcal{R} = J \times [n]$ we use a simplified notation: for $J \subset [n]$

$$A_J := A_{[n] \times J}.$$

Given a finite set S , by $|S|$ we denote its cardinality. The standard inner product in \mathbb{R}^n shall be denoted by $\langle \cdot, \cdot \rangle$. Given $p \in [1, \infty]$, $\|\cdot\|_p$ is the standard ℓ_p^n -norm in \mathbb{R}^n . Also, $\|\cdot\|_{\psi_2}$ denotes sub-gaussian norm of a random variable and $\|\cdot\|_{\psi_1}$ – sub-exponential norm (see also in Section 3.3).

3.2. Operator norm via ℓ_1 norm of rows and columns. The following simple result states that the operator norm of any matrix is dominated by the ℓ_1 norms of rows and columns.

Lemma 3.1. *For any $m \times k$ matrix A , we have*

$$\|A\| \leq \left(\max_i \|A_i\|_1 \cdot \max_j \|A^j\|_1 \right)^{1/2}$$

where A_i and A^j denote the rows and columns of A .

Proof. Recall that the operator norm can be computed as a maximum of the quadratic form:

$$\|A\| = \sup_{\|x\|_2 = \|y\|_2 = 1} |x^\top A y|.$$

Fix unit vectors x and y and express

$$\begin{aligned}
|x^\top A y| &= \left| \sum_{i,j} x_i A_{ij} y_j \right| \\
&\leq \sum_{i,j} \left(|x_i| \sqrt{|A_{ij}|} \right) \left(\sqrt{|A_{ij}|} |y_j| \right) \quad (\text{by triangle inequality}) \\
&\leq \left(\sum_{i,j} x_i^2 |A_{ij}| \right)^{1/2} \left(\sum_{i,j} |A_{ij}| y_j^2 \right)^{1/2} \quad (\text{by Cauchy-Schwarz}) \\
&= \left(\sum_i x_i^2 \|A_i\|_1 \right)^{1/2} \left(\sum_j \|A_j\|_1 y_j^2 \right)^{1/2} \\
&\leq \max_i \|A_i\|_1^{1/2} \cdot \max_j \|A_j\|_1^{1/2} \quad (\text{since } \|x\|_2 = \|y\|_2 = 1).
\end{aligned}$$

Taking the maximum over all unit vectors x and y , we complete the proof. \square

3.3. Concentration. A standard way to get some desired estimate on a random variable X *with high probability* is to get this estimate for $\mathbb{E} X$ first, and then argue that X *concentrates* around its expectation. In this case X usually stays close to $\mathbb{E} X$, and therefore satisfies a close estimate.

In this paper we make use of good concentration properties of the sums of sub-gaussian (and sub-exponential) random variables, that is, such that grow not faster than standard normal (respectively, exponential) random variables. Recall that by definition a random variable Y is called *sub-gaussian* if its moments satisfy

$$\mathbb{E} \exp(Y^2/M_2^2) \leq e,$$

for some number $M_2 > 0$. The minimal number M_2 is called the sub-gaussian moment of X , denoted as $\|Y\|_{\psi_2}$. Analogously, a random variable is called *sub-exponential* if

$$\mathbb{E} \exp(Y/M_1) \leq e,$$

for some number $M_1 > 0$. The minimal number M_1 is called the sub-exponential moment of Y , denoted as $\|Y\|_{\psi_1}$.

The class of sub-gaussian random variables contains standard normal, Bernoulli, and generally all bounded random variables. The class of sub-exponential random variables is exactly the class of squares of sub-gaussians. See [18] for more information and statements of standard concentration inequalities.

Also we will need a concentration inequality for random permutations from [13].

Lemma 3.2 (Concentration for random permutations). *Consider arbitrary vectors $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ and $x \in \{-1, 1\}^n$. Let $\pi : [n] \rightarrow [n]$ denote a*

random permutation chosen uniformly from the symmetric group S_n . Then the random sum

$$S := \sum_{i=1}^n a_i x_{\pi(i)}$$

is sub-gaussian, and

$$\|S - \mathbb{E} S\|_{\psi_2} \leq C \|a\|_2.$$

The same inequality holds for the sum $S' = \sum_{i=1}^n a_{\pi(i)} x_i$ as well, since it has the same distribution as S .

3.4. Discretization. The following lemma allows us to approximate a general continuous random variable by a sum of independent, scaled Bernoulli random variables. This lemma was originally proved in [13]. Here we give a proof for completeness, and then discuss some particular cases needed for the proof of Theorem 1.1.

Lemma 3.3 (Discretization). *Consider a non-negative, continuous random variable X . There exists a non-negative random variable X' satisfying the following.*

1. $\mathbb{E} X' \leq 4 \mathbb{E} X$.
2. X' stochastically dominates X , i.e.

$$\mathbb{P}\{X' \geq t\} \geq \mathbb{P}\{X \geq t\} \quad \text{for all } t \geq 0.$$

3. X' is a sum of scaled, independent Bernoulli random variables:

$$X' = \sum_{k=0}^{\infty} q_k \xi_k \tag{3.1}$$

where q_k are non-negative numbers and ξ_k are independent $\text{Ber}(2^{-k})$ random variables.

Proof. Set the values q_k to be the quantiles of the distribution of X :

$$q_k := \min \left\{ t \geq 0 : \mathbb{P}\{X \geq t\} = 2^{-k-1} \right\}, \quad k = 0, 1, 2, \dots$$

(These values are well defined since the cumulative distribution function of X is continuous by assumption.) By definition, (q_k) is an increasing sequence. Define X' by (3.1).

To check part 1, note that by definition,

$$\mathbb{E} X' = \sum_{k=0}^{\infty} q_k \mathbb{E} \xi_k = \sum_{k=0}^{\infty} q_k 2^{-k}. \tag{3.2}$$

To lower bound $\mathbb{E} X$, let us decompose X according to the values it can take. This gives

$$X \geq \sum_{k=0}^{\infty} X \mathbf{1}_{\{X \in [q_k, q_{k+1})\}} \geq \sum_{k=0}^{\infty} q_k \mathbf{1}_{\{X \in [q_k, q_{k+1})\}}$$

almost surely. Taking expectation of both sides, we obtain

$$\mathbb{E} X \geq \sum_{k=0}^{\infty} q_k \mathbb{P} \{X \in [q_k, q_{k+1})\}.$$

Now, using the definition of q_k , we have

$$\mathbb{P} \{X \in [q_k, q_{k+1})\} = \mathbb{P} \{X \geq q_k\} - \mathbb{P} \{X \geq q_{k+1}\} = 2^{-k-1} - 2^{-k-2} = 2^{-k-2}.$$

This yields

$$\mathbb{E} X \geq \sum_{k=0}^{\infty} q_k 2^{-k-2}. \quad (3.3)$$

Comparing (3.2) with (3.3), we conclude that $\mathbb{E} X' \leq 4 \mathbb{E} X$, which proves part 1 of the lemma.

Let us prove part 2. If $t \in [q_k, q_{k+1})$ for some $k = 0, 1, 2, \dots$, then using the definitions of X' and q_k we obtain

$$\begin{aligned} \mathbb{P} \{X' \geq t\} &\geq \mathbb{P} \{X' \geq q_{k+1}\} \geq \mathbb{P} \{\xi_{k+1} = 1\} = 2^{-k-1} \\ &= \mathbb{P} \{X \geq q_k\} \geq \mathbb{P} \{X \geq t\}, \end{aligned}$$

as required.

It remains to check the domination inequality when t is outside the range $[q_0, q_{\infty})$ where $q_{\infty} := \lim_{k \rightarrow \infty} q_k \in \mathbb{R}_+ \cup \{\infty\}$. If $t < q_0$, we have

$$\mathbb{P} \{X' \geq t\} \geq \mathbb{P} \{X' \geq q_0\} \geq \mathbb{P} \{\xi_0 = 1\} = 1,$$

and the inequality in part 2 follows. If $t \geq q_{\infty}$ then, using the continuity of the cumulative distribution of X , we obtain

$$\mathbb{P} \{X \geq t\} \leq \mathbb{P} \{X \geq q_{\infty}\} = \lim_{k \rightarrow \infty} \mathbb{P} \{X \geq q_k\} = \lim_{k \rightarrow \infty} 2^{-k-1} = 0,$$

and the inequality in part 2 follows again. The proof is complete. \square

Remark 3.4 (Bounded random variables). Suppose $X \leq M$ almost surely. Then, in the second part of the conclusion of Lemma 3.3, X can be represented as a *finite* sum

$$X' := \sum_{k=0}^{\kappa} q_k \xi_k$$

where q_k are non-negative numbers, $q_k \in [0, M]$, and ξ_k are independent $\text{Ber}(p_k)$ random variables. Here $p_k = 2^{-k} \geq 1/M$ for $k < \kappa$ and $p_{\kappa} = 1/M$.

Remark 3.5 (Coupling). Stochastic dominance of X' over X in Lemma 3.3 implies that one can realize the random variables X and X' on the same probability space so that

$$X' \geq X \quad \text{almost surely.}$$

(See, for example, [19, Section 4.3]).

Moreover, in the same way we can construct a majorizing collection for any collection of independent random variables. In particular, we can do it for all entries of the matrix A at once.

4. DAMPING A SUM OF INDEPENDENT RANDOM VARIABLES

Let X_1, \dots, X_n be non-negative i.i.d. random variables with $\mathbb{E} X_i \leq 1$. The linearity of expectation gives the trivial bound

$$\mathbb{E} \sum_{i=1}^n X_i \leq n.$$

Here we will be interested in a stronger result – that the sum be $O(n)$ *almost surely* instead of in expectation. To do this, we will be looking for random weights

$$W_1, \dots, W_n \in [0, 1]$$

that make the “damped” sum satisfy

$$\sum_{j=1}^n W_j X_j = O(n) \quad \text{almost surely.}$$

To make the damping as gentle as possible, we are looking for largest possible weights W_i , hopefully very close to 1.

4.1. Damping one random variable. To get started, let us consider the simple case where $n = 1$ and try to damp one random variable.

Lemma 4.1 (Damping a random variable). *Let X be a random variable such that*

$$X \geq 0 \quad \text{and} \quad \mathbb{E} X \leq 1.$$

Let $\varepsilon \in (0, 1)$. There exists a random variable W taking values in $[0, 1]$ and such that

$$XW \leq \varepsilon^{-1} \quad \text{almost surely;} \tag{4.1}$$

$$1 \leq \mathbb{E} W^{-1} \leq 1 + \varepsilon. \tag{4.2}$$

Proof. Fix a level $L \geq 1$ whose value we will choose later, and define

$$W := \min(1, L/X).$$

To check (4.1), we have

$$XW = \min(X, L) \leq L \quad \text{almost surely.}$$

Next, the lower bound in (4.2) holds trivially since $W \leq 1$. For the upper bound, we have

$$\mathbb{E} W^{-1} = \mathbb{E} \max(1, X/L) \leq \mathbb{E}(1 + X/L) \leq 1 + \frac{1}{L},$$

where we used the assumption that $\mathbb{E} X \leq 1$. Setting $L = \varepsilon^{-1}$ completes the proof. \square

4.2. Damping a sum of random variables. Now let us address the damping problem for general number n of random variables, which we described in the beginning of this section. Applying Lemma 4.1 for each random variable X_i , we get weights W_i such that

$$\sum_{j=1}^n W_j X_j \leq \varepsilon^{-1} n \quad \text{almost surely;}$$

$$1 \leq \mathbb{E} \left(\prod_{j=1}^n W_j \right)^{-1} \leq (1 + \varepsilon)^n = 1 + O(\varepsilon n)$$

for small ε . We will now considerably improve both these bounds, making only one mild extra assumption that $X_i = O(n)$ almost surely.

Theorem 4.2 (Damping a sum of random variables). *Let X_1, \dots, X_n be i.i.d. random variables such that*

$$0 \leq X_j \leq Kn \quad \text{and} \quad \mathbb{E} X_j \leq 1$$

for some $K \geq 1$. Let $\varepsilon \in (0, 1/2)$. There exist random variables W_1, \dots, W_n taking values in $[0, 1]$ and such that

$$\sum_{j=1}^n W_j X_j \leq CK \log(\varepsilon^{-1}) \cdot n \quad \text{almost surely;} \quad (4.3)$$

$$1 \leq \mathbb{E} \left(\prod_{j=1}^n W_j \right)^{-1} \leq 1 + \varepsilon. \quad (4.4)$$

Remark 4.3. Improvement in the order of n in (4.4) does not require an extra boundedness assumption, and it was done in previous work [13]. We employ the same ideas as in [13, Lemma 3.3] and obtain better (logarithmic) dependence on ε in (4.3) in trade of the additional assumption mentioned.

Proof. Step 1: Bernoulli distribution. Let us first prove the theorem in the partial case where X_j are scaled Bernoulli random variables. Assume that X_j can take values q and 0, and

$$\mathbb{P} \{X_j = q\} = p \geq \frac{1}{Kn}. \quad (4.5)$$

Let ν denote the (random) number of nonzero X_j 's:

$$\nu := |\{j : X_j \neq 0\}|, \quad \text{then} \quad \mathbb{E} \nu = pn.$$

Here is how we will define the weights W_j . If $X_j = 0$ then clearly there is no need to damp X_j so put $W_j = 1$. The same applies if the number ν of non-zero X_j 's does not significantly exceed its expectation pn . Otherwise we damp all terms by the same amount $W_j \sim pn/\nu$. Formally, we fix some parameter $L = L(K, \varepsilon)$ whose value we will determine later, and set

$$W_j := \begin{cases} 1, & \text{if } \nu \leq Lpn \text{ or } X_j = 0 \\ Lpn/\nu, & \text{if } \nu > Lpn \text{ and } X_j \neq 0. \end{cases}$$

Let us check (4.3). In the event when $\nu \leq Lpn$, we have

$$\sum_{j=1}^n W_j X_j = \sum_{j=1}^{\nu} 1 \cdot q = q\nu \leq qLpn = Ln \cdot \mathbb{E} X_1.$$

And in the event when $\nu > Lpn$, we have

$$\sum_{j=1}^n W_j X_j = \sum_{j=1}^{\nu} \frac{Lpn}{\nu} \cdot q = Lpnq = Ln \cdot \mathbb{E} X_1.$$

as before. Thus, we showed that

$$\sum_{j=1}^n W_j X_j \leq Ln \cdot \mathbb{E} X_1 \leq Ln \quad \text{almost surely.} \quad (4.6)$$

Let us now check (4.4). Since the lower bound is trivial, we will only have to check the upper bound. We will again split the calculation into two cases based on the size of ν . If $\nu \leq Lpn$ then all $W_j = 1$, so we trivially get

$$E_- := \mathbb{E} \left(\prod_{j=1}^n W_j \right)^{-1} \mathbb{1}_{\{\nu \leq Lpn\}} \leq 1.$$

If $\nu > Lpn$, then the definition of W_j gives

$$\begin{aligned} E_+ &:= \mathbb{E} \left(\prod_{j=1}^n W_j \right)^{-1} \mathbb{1}_{\{\nu > Lpn\}} = \mathbb{E} \left(\frac{\nu}{Lpn} \right)^{\nu} \mathbb{1}_{\{\nu > Lpn\}} \\ &= \sum_{k=\lceil Lpn \rceil + 1}^n \left(\frac{k}{Lpn} \right)^k \mathbb{P} \{ \nu = k \}. \end{aligned}$$

Since $\nu \sim \text{Binom}(n, p)$, we have

$$\mathbb{P} \{ \nu = k \} = \binom{n}{k} p^k \leq \left(\frac{enp}{k} \right)^k,$$

using a standard consequence of Stirling's approximation. Thus

$$E_+ \leq \sum_{k=\lceil Lpn \rceil + 1}^n \left(\frac{e}{L} \right)^k \leq \left(\frac{e}{L} \right)^{Lpn},$$

provided that $L \geq 10$. Thus we showed that

$$\mathbb{E} \left(\prod_{j=1}^n W_j \right)^{-1} \leq E_- + E_+ \leq 1 + \left(\frac{e}{L} \right)^{Lpn} \leq 1 + \left(\frac{e}{L} \right)^{L/K} \quad (4.7)$$

where in the last step we used the assumption that $p \geq 1/Kn$ that we made in (4.5).

Now that we have the bounds (4.6) and (4.7), it is enough to choose

$$L := CK \log \left(\frac{1}{\varepsilon} \right)$$

which implies that $E \leq 1 + \varepsilon$. The proof for the Bernoulli distribution is complete.

Step 2. General distribution. Let us now prove the theorem in full generality. First we discretize the distribution of X_j using Lemma 3.3. This result requires X_j be continuous, which can be arranged by a standard approximation argument. For example, we can add a small Gaussian independent component to X_j and then let the variance of this component go to zero. Taking into account Remarks 3.4 and 3.5, we obtain independent, non-negative random variables X'_j that satisfy $\mathbb{E} X'_j \leq 4$ and such that

$$X_j \leq X'_j = \sum_{k=1}^{\kappa} X_{jk}.$$

Here X_{jk} are independent random variables; each X_{jk} can take values q_k and 0, and

$$\mathbb{P}\{X_{jk} = q_k\} = p_k$$

with

$$p_k = 2^{-k} \geq \frac{1}{Kn} \text{ for } k < \kappa, \quad p_{\kappa} = \frac{1}{Kn}. \quad (4.8)$$

The argument will be similar to step 1 of the proof. For each level k we let ν_k denote number of non-zero X_{jk} 's:

$$\nu_k := |\{j : X_{jk} \neq 0\}|, \quad \text{then} \quad \mathbb{E} \nu = p_k n.$$

Again, for each level k define the weights W_{jk} like in step 1:

$$W_{jk} := \begin{cases} 1, & \text{if } \nu_k \leq Lp_k n \text{ or } X_{jk} = 0 \\ Lp_k n / \nu_k, & \text{if } \nu_k > Lp_k n \text{ and } X_{jk} \neq 0. \end{cases}$$

Then we set

$$W_j := \prod_{k=1}^{\kappa} W_{jk}, \quad j = 1, \dots, n.$$

Let us check (4.3). We have

$$\sum_{j=1}^n W_j X_j \leq \sum_{j=1}^n W_j X'_j = \sum_{j=1}^n \sum_{k=1}^{\kappa} W_j X_{jk} \leq \sum_{k=1}^{\kappa} \sum_{j=1}^n W_{jk} X_{jk}, \quad (4.9)$$

since $W_j \leq W_{jk}$ by construction. Now, for each level k , we can use step 1 of the proof, where we showed in (4.6) that

$$\sum_{j=1}^n W_{jk} X_{jk} \leq Ln \cdot \mathbb{E} X_{1k}.$$

Substituting into (4.9), we obtain

$$\sum_{j=1}^n W_j X_j \leq Ln \cdot \sum_{k=1}^{\kappa} \mathbb{E} X_{1k} = Ln \cdot \mathbb{E} X'_1 \leq 5Ln \quad (4.10)$$

by construction.

Let us now check (4.4). The lower bound is trivial, and we will only have to check the upper bound. For each level k , we can use step 1 of the proof, where we showed in (4.7) that

$$\mathbb{E} \left(\prod_{j=1}^n W_{jk} \right)^{-1} \leq 1 + \left(\frac{e}{L} \right)^{Lp_k n} \leq 1 + e^{-Lp_k n},$$

which is true as long as $L \geq 10$. Then, by construction we have

$$\begin{aligned} \mathbb{E} \left(\prod_{j=1}^n W_j \right)^{-1} &= \mathbb{E} \prod_{k=1}^{\kappa} \left(\prod_{j=1}^n W_{jk} \right)^{-1} \\ &= \prod_{k=1}^{\kappa} \mathbb{E} \left(\prod_{j=1}^n W_{jk} \right)^{-1} \quad (\text{by independence}) \\ &\leq \prod_{k=1}^{\kappa} (1 + e^{-Lp_k n}) \leq \exp \left(\sum_{k=1}^{\kappa} e^{-Lp_k n} \right) \end{aligned}$$

where in the last step we used the inequality $1 + x \leq e^x$. Recall from (4.8) that the exponents p_k form a decreasing geometric progression with values 2^{-k} until the last (smallest) term of order $1/Kn$. So this last term dominates the sum $\sum_{k=1}^{\kappa} e^{-Lp_k n}$, and we obtain

$$\mathbb{E} \left(\prod_{j=1}^n W_j \right)^{-1} \leq \exp(2e^{-L/2K}). \quad (4.11)$$

Now that we have the bounds (4.10) and (4.11), it is enough to choose

$$A_{ij}L := C_{4.2}K \log \left(\frac{1}{\varepsilon} \right)$$

with $C_{4.2} \geq 6K$ and the right hand side of (4.11) will be bounded by

$$\exp(2\varepsilon^3) \leq \exp(\varepsilon/2) \leq 1 + \varepsilon,$$

as claimed. The proof of the theorem is complete. \square

5. THE $2 \rightarrow \infty$ NORM OF RANDOM MATRICES

In this section we prove Theorem 1.1 under the additional assumption that all entries A_{ij} of A are not too large. Specifically, let us assume that

$$|A_{ij}| \leq \frac{\sqrt{n}}{2} \quad \text{almost surely.} \quad (5.1)$$

Lemma 5.1 (Bounding $2 \rightarrow \infty$ norm by removing a few columns). *Consider an $n \times n$ random matrix A with i.i.d. entries A_{ij} which have mean zero and unit variance and satisfy (5.1). Let $\varepsilon \in (0, 1/2]$. Then with probability at least $1 - \exp(-\varepsilon n)$, there exists a subset $J \in [n]$ with cardinality $|J| \leq \varepsilon n$ such that*

$$\|A_{J^c}\|_{2 \rightarrow \infty} \leq C\sqrt{\ln \varepsilon^{-1}} \cdot \sqrt{n}.$$

Proof. We apply Theorem 4.2 for the squares of the elements in each row of A , i.e. for the random variables $(a_{i1}^2, \dots, a_{in}^2)$. This gives us random weights $W_{ij} \in [0, 1]$ which satisfy for each $i \in [n]$ that

$$\sum_{j=1}^n W_{ij} A_{ij}^2 \leq C \log(\varepsilon^{-1})n \quad \text{a.s.}; \quad \mathbb{E} \left(\prod_{j=1}^n W_{ij} \right)^{-1} \leq \exp(\varepsilon).$$

To make the same system of weights work for all rows, we define

$$V_j := \prod_{i=1}^n W_{ij} \in [0, 1], \quad j \in [n].$$

Then obviously $V_j \leq W_{ij}$ for every i , and so

$$\sum_{j=1}^n V_j A_{ij}^2 \leq C \log(\varepsilon^{-1})n \quad \forall i \quad \text{a.s.}; \quad \mathbb{E} \left(\prod_{j=1}^n V_j \right)^{-1} \leq \exp(\varepsilon n). \quad (5.2)$$

We will remove from A the columns whose weights V_j are too small, namely those in

$$J := \{j \in [n] : V_j < e^{-2}\}.$$

Let us first check that

$$|J| \leq \varepsilon n \quad \text{with probability at least } 1 - \exp(-\varepsilon n), \quad (5.3)$$

as we claimed in the lemma. Indeed, if $|J| > \varepsilon n$ then using that all $V_j \in [0, 1]$ we have

$$Z := \prod_{j=1}^n V_j \leq \prod_{j \in J} V_j < e^{-2\varepsilon n}.$$

But the probability of this event can be bounded by Markov's inequality:

$$\mathbb{P} \{Z < e^{-2\varepsilon n}\} = \mathbb{P} \{Z^{-1} > e^{2\varepsilon n}\} \leq e^{-2\varepsilon n} \mathbb{E} Z^{-1} \leq e^{-\varepsilon n},$$

where in the last bound we used (5.2). This proves (5.3).

It remains to check that all rows B_i of the matrix $B = A_{[n] \times J_0^c}$ are bounded as claimed. We have

$$\begin{aligned} \|B_i\|_2^2 &= \sum_{j \in J^c} A_{ij}^2 \leq e^2 \sum_{j \in J^c} V_j A_{ij}^2 \quad (\text{by definition of } J) \\ &\leq e^2 \sum_{j=1}^n V_j A_{ij}^2 \quad (\text{since all } V_j \leq 1) \\ &\leq e^2 C \log(\varepsilon^{-1})n \quad (\text{by (5.2)}). \end{aligned}$$

Taking the square root of both sides completes the proof. \square

6. FROM $2 \rightarrow \infty$ NORM TO $\infty \rightarrow 2$ NORM

In this section we will control the $\infty \rightarrow 2$ norm of a random matrix. Our first task is to bound the $\infty \rightarrow 2$ norm by the simpler $2 \rightarrow \infty$ norm. There are two ways to do this, both of them going back to [13]. The resulting comparison inequalities are interesting in their own right; we state them in Lemmas 6.1 and 6.3. The ultimate result of this section is Theorem 6.6, which gives an optimal bound $O(n)$ on the $\infty \rightarrow 2$ norm of a random matrix after removing a small fraction of columns.

6.1. Using random signs. The first method is based on flipping the signs of the entries independently at random. Here is the main result of this section.

Lemma 6.1 (From $2 \rightarrow \infty$ to $\infty \rightarrow 2$). *Let A be an $n \times n$ random matrix whose entries are independent, mean zero random variables. Then*

$$\mathbb{E} \|A\|_{\infty \rightarrow 2} \leq C\sqrt{n} \cdot \mathbb{E} \|A\|_{2 \rightarrow \infty}.$$

Proof. Let ε_{ij} be independent Rademacher random variables (which are also independent of A) and consider the random matrix

$$\tilde{A} := (\varepsilon_{ij} A_{ij}).$$

A basic symmetrization inequality (see [11, Lemma 6.3]) yields

$$\mathbb{E} \|A\|_{\infty \rightarrow 2} \leq 2 \mathbb{E} \|\tilde{A}\|_{\infty \rightarrow 2}.$$

Condition on A ; the randomness now rests in the random signs (ε_{ij}) only. It suffices to show that the conditional expectation satisfies

$$\mathbb{E} \|\tilde{A}\|_{\infty \rightarrow 2} \lesssim \sqrt{n} \cdot \|A\|_{2 \rightarrow \infty}. \quad (6.1)$$

Recalling (2.2), we have

$$\|\tilde{A}\|_{\infty \rightarrow 2} = \max_{x \in \{-1, 1\}^n} \|\tilde{A}x\|_2. \quad (6.2)$$

According to the matrix-vector multiplication, we can express $\|\tilde{A}x\|^2$ as a sum of independent random variables

$$\|\tilde{A}x\|_2^2 = \sum_{i=1}^n \xi_i^2 \quad \text{where} \quad \xi_i := \langle \tilde{A}_i, x \rangle = \sum_{j=1}^n \varepsilon_{ij} A_{ij} x_j.$$

Fix $x \in \{-1, 1\}^n$. Using independence and (2.1), we get

$$\mathbb{E} \xi_i^2 = \sum_{j=1}^n (A_{ij} x_{ij})^2 = \sum_{j=1}^n A_{ij}^2 \leq \|A\|_{2 \rightarrow \infty}^2,$$

so

$$\mathbb{E} \sum_{i=1}^n \xi_i^2 \leq n \|A\|_{2 \rightarrow \infty}^2. \quad (6.3)$$

Moreover, the standard concentration results ([18, Lemma 5.9]) show that each ξ_i is a sub-gaussian random variable, and we have

$$\|\xi_i\|_{\psi_2}^2 = \left\| \sum_{j=1}^n \varepsilon_{ij} A_{ij} x_j \right\|_{\psi_2}^2 \lesssim \sum_{j=1}^n (A_{ij} x_{ij})^2 \leq \|A\|_{2 \rightarrow \infty}^2.$$

Thus ξ_i^2 is a sub-exponential random variable (see [18, Lemma 5.9]) and

$$\|\xi_i^2\|_{\psi_1} \lesssim \|\xi_i\|_{\psi_2}^2 \lesssim \|A\|_{2 \rightarrow \infty}^2. \quad (6.4)$$

Applying Bernstein's concentration inequality [18, Corollary 5.17] together with (6.3) and (6.4), we obtain

$$\mathbb{P} \left\{ \sum_{i=1}^n \xi_i^2 \geq n\|A\|_{2 \rightarrow \infty}^2 + tn\|A\|_{2 \rightarrow \infty}^2 \right\} \leq \exp(-ctn)$$

for all $t \geq 1$. Thus we obtained a bound on $\|\tilde{A}x\|_2^2 = \sum_{i=1}^n \xi_i^2$. It remains to recall (6.2) and take a union bound over $x \in \{-1, 1\}^n$. It follows that the inequality

$$\|\tilde{A}\|_{\infty \rightarrow 2}^2 \leq (1+t)n\|A\|_{2 \rightarrow \infty}^2$$

holds with probability at least

$$1 - 2^n \exp(-ctn) \geq 1 - \exp(-(1-ct)n),$$

where $t \geq 1$ is arbitrary. Integration of these tails implies (6.1). \square

We will need a minor variation of Lemma 6.1 that can be applied even when some of the columns of A are removed.

Lemma 6.2 (From $2 \rightarrow \infty$ to $\infty \rightarrow 2$ for symmetric distributions). *Let A be an $n \times n$ random matrix whose entries are independent, symmetric random variables. Let $J \subset [n]$ be a random subset, which is independent of the signs of the entries of A . Then*

$$\|A_J\|_{\infty \rightarrow 2} \leq C\sqrt{n}\|A_J\|_{2 \rightarrow \infty}$$

with probability at least $1 - e^{-n}$.

Proof. It is quite straightforward to check this result by modifying the proof of Lemma 6.1. By the symmetry assumption, the matrix $\tilde{A} := (\varepsilon_{ij} A_{ij})$ has the same distribution as A . Conditioning on A and J leaves all randomness with the signs (ε_{ij}) , as before. Then we repeat the rest of the proof of Lemma 6.1 for the submatrix A_J . In the end, we choose t to be a large absolute constant to complete the proof. \square

So, the only part of Lemma 6.1 that does not work for a matrix with removed columns is the symmetrization part. In the following two sections we will develop the tools to overcome the extra symmetry assumption we have to add in Lemma 6.2.

6.2. Using random permutations. We just showed how to convert an $\infty \rightarrow 2$ bound to a $2 \rightarrow \infty$ bound for random matrices by using random signs. Alternatively, one can use random permutations for the same purpose, and obtain the following bound.

Lemma 6.3 (From $2 \rightarrow \infty$ to $\infty \rightarrow 2$). *Let A be an $n \times n$ random matrix with i.i.d. entries. Then*

$$\mathbb{E} \|A\|_{\infty \rightarrow 2} \leq C\sqrt{n} \cdot \mathbb{E} \|A\|_{2 \rightarrow \infty} + C \mathbb{E} \|A\mathbf{1}\|_2,$$

where $\mathbf{1} = (1, 1, \dots, 1)$ denotes the vector whose all coordinates equal 1.

Before we turn to the proof, note that the only difference between Lemmas 6.1 and 6.3 is the term $\mathbb{E} \|A\mathbf{1}\|_2$. It makes its appearance since there is no mean zero assumption on the entries. This term is usually quite innocent. Note also that (2.2) trivially implies that

$$\mathbb{E} \|A\|_{\infty \rightarrow 2} \geq \mathbb{E} \|A\mathbf{1}\|_2,$$

so we have to control this term anyway.

Proof. Let us apply a random independent permutation π_i to the elements of each row of A . The resulting matrix \tilde{A} has the same distribution of A due to the i.i.d. assumption. Condition on A ; the randomness now rests in the random permutations π_i only. It suffices to show that the conditional expectation satisfies

$$\mathbb{E} \|\tilde{A}\|_{\infty \rightarrow 2} \leq C\sqrt{n} \cdot \|A\|_{2 \rightarrow \infty} + C\|A\mathbf{1}\|_2, \quad (6.5)$$

Similarly to the proof of Lemma 6.1, we express $\|\tilde{A}x\|^2$ as a sum of independent random variables

$$\|\tilde{A}x\|_2^2 = \sum_{i=1}^n \xi_i^2 \quad \text{where} \quad \xi_i := \langle \tilde{A}_i, x \rangle = \sum_{j=1}^n A_{i, \pi_i(j)} x_j. \quad (6.6)$$

The concentration inequality for random permutations (Lemma 3.2) states that each ξ_i is a sub-gaussian random variable, and we have

$$\|\xi_i - \mathbb{E} \xi_i\|_{\psi_2} \lesssim \|\tilde{A}_i\|_2 \leq \|A\|_{2 \rightarrow \infty}.$$

Just like in the proof of Lemma 6.1, this implies that

$$\|(\xi_i - \mathbb{E} \xi_i)^2\|_{\psi_1} \lesssim \|A\|_{2 \rightarrow \infty}^2.$$

Since the expectation is bounded by the ψ_1 norm (see e.g. [18, Definition 5.13]), we conclude that

$$\mathbb{E}(\xi_i - \mathbb{E} \xi_i)^2 \lesssim \|(\xi_i - \mathbb{E} \xi_i)^2\|_{\psi_1} \lesssim \|A\|_{2 \rightarrow \infty}^2$$

and thus

$$\mathbb{E} \sum_{i=1}^n (\xi_i - \mathbb{E} \xi_i)^2 \lesssim n \|A\|_{2 \rightarrow \infty}^2.$$

Applying Bernstein's inequality like in Lemma 6.1, we find that

$$\mathbb{P} \left\{ \sum_{i=1}^n (\xi_i - \mathbb{E} \xi_i)^2 \geq n \|A\|_{2 \rightarrow \infty}^2 + tn \|A\|_{2 \rightarrow \infty}^2 \right\} \leq \exp[-c \min(t^2, t)n]$$

for all $t \geq 0$. Thus, for any $t \geq 1$ we have with probability at least $1 - \exp(-tn)$ that

$$\sum_{i=1}^n (\xi_i - \mathbb{E} \xi_i)^2 \leq (1+t)n \|A\|_{2 \rightarrow \infty}^2. \quad (6.7)$$

From (6.6) we see that we are almost done; we just need to remove $\mathbb{E} \xi_i$ from our bound. To this end, note that

$$\|\tilde{A}x\|_2^2 = \sum_{i=1}^n \xi_i^2 \leq 2 \sum_{i=1}^n (\xi_i - \mathbb{E} \xi_i)^2 + 2 \sum_{i=1}^n (\mathbb{E} \xi_i)^2. \quad (6.8)$$

We have already bounded the first sum. As for the second one, the definition of ξ in (6.6) yields

$$\mathbb{E} \xi_i = \frac{2m-n}{n} \sum_{j=1}^n A_{ij} = \frac{2m-n}{n} \langle A_i, \mathbf{1} \rangle$$

where m denotes the number of ones in x_j and A_i^\top is the i -th row of A . Thus

$$\sum_{i=1}^n (\mathbb{E} \xi_i)^2 = \left(\frac{2m-n}{n} \right)^2 \sum_{i=1}^n \langle A_i, \mathbf{1} \rangle^2 \leq \|A\mathbf{1}\|_2^2.$$

We substitute this and (6.7) into (6.8) and obtain that for any $t \geq 1$,

$$\|\tilde{A}x\|_2^2 \leq 2(1+t)n \|A\|_{2 \rightarrow \infty}^2 + 2\|A\mathbf{1}\|_2^2$$

with probability at least $1 - \exp(-tn)$.

It remains to recall (6.2) and take a union bound over $x \in \{-1, 1\}^n$. It follows that the inequality

$$\|\tilde{A}\|_{\infty \rightarrow 2}^2 \leq 2(1+t)n \|A\|_{2 \rightarrow \infty}^2 + 2\|A\mathbf{1}\|_2^2 \quad (6.9)$$

holds with probability at least

$$1 - 2^n \exp(-ctn) \geq 1 - \exp[(1-ct)n]$$

where $t \geq 1$ is arbitrary. Integration of these tails implies (6.5). \square

It is worthwhile to mention a high-probability version of Lemma 6.3.

Lemma 6.4 (From $2 \rightarrow \infty$ to $\infty \rightarrow 2$ with high probability). *Let A be an $n \times n$ random matrix with i.i.d. entries. Then with probability at least $1 - e^{-n}$ we have*

$$\|A\|_{\infty \rightarrow 2} \leq C\sqrt{n} \cdot \mathbb{E} \|A\|_{2 \rightarrow \infty} + C \mathbb{E} \|A\mathbf{1}\|_2,$$

where $\mathbf{1} = (1, 1, \dots, 1)$ denotes the vector whose all coordinates equal 1.

Proof. At the end of the proof of Lemma 6.3, we obtained inequality (6.9) which states (for large constant t) that

$$\|\tilde{A}\|_{\infty \rightarrow 2} \leq C\sqrt{n} \cdot \|A\|_{2 \rightarrow \infty} + C\|A\mathbf{1}\|_2$$

with probability at least $1 - e^{-n}$. Note that

$$\|A\|_{2 \rightarrow \infty} = \|\tilde{A}\|_{2 \rightarrow \infty} \quad \text{and} \quad \|A\mathbf{1}\|_2 = \|\tilde{A}\mathbf{1}\|_2$$

deterministically. Indeed, it is easy to check that permutations of the elements of the rows of A do not affect these two quantities. It follows that

$$\|\tilde{A}\|_{\infty \rightarrow 2} \leq C\sqrt{n} \cdot \|\tilde{A}\|_{2 \rightarrow \infty} + C\|\tilde{A}\mathbf{1}\|_2$$

with probability at least $1 - e^{-n}$. It remains to note that \tilde{A} has the same distribution as A . \square

6.3. Bounding $2 \rightarrow \infty$ and $\infty \rightarrow 2$ norms with tiny probability.

Recall from Section 2.2 that ideally, we would want

$$\|A\|_{2 \rightarrow \infty} \lesssim \sqrt{n} \quad \text{and} \quad \|A\|_{\infty \rightarrow 2} \lesssim n$$

with high probability. But this is too good to be true in our situation, where we assume only two moments for the entries of A . Nevertheless, we will now show that these bounds still hold, albeit with exponentially small probability.

Lemma 6.5 ($2 \rightarrow \infty$ and $\infty \rightarrow 2$ norms with tiny probability). *Let A be an $n \times n$ random matrix whose entries are i.i.d. random variables with mean zero and unit variance. Let $\delta \in (0, 1/2)$. Then*

$$\|A\|_{2 \rightarrow \infty} \leq 2\delta^{-1}\sqrt{n} \quad \text{and} \quad \|A\|_{\infty \rightarrow 2} \leq C\delta^{-1}n \quad (6.10)$$

with probability at least $\frac{1}{2} \exp(-\delta^2 n)$.

Proof. We will first bound below the probability of the event

$$\mathcal{E} := \left\{ \|A\|_{2 \rightarrow \infty} \leq 2\delta^{-1}\sqrt{n} \text{ and } \|\tilde{A}\mathbf{1}\|_2 \leq 2\delta^{-1}n \right\}$$

and then use Lemma 6.4 to control $\|A\|_{\infty \rightarrow 2}$.

Recall from (2.1) that

$$\|A\|_{2 \rightarrow \infty} = \max_{i \in [n]} \|A_i\|_2 \quad \text{and} \quad \|\tilde{A}\mathbf{1}\|_2^2 = \sum_{i=1}^n \langle A_i, \mathbf{1} \rangle^2$$

where A_i^\top denote the rows of A . Thus $\mathcal{E} \subset \bigcap_{i=1}^n \mathcal{E}_i$ where

$$\mathcal{E}_i := \left\{ \|A_i\|_2 \leq 2\delta^{-1}\sqrt{n} \text{ and } |\langle A_i, \mathbf{1} \rangle| \leq 2\delta^{-1}\sqrt{n} \right\}$$

are independent events. This reduces the problem to bounding the probability of each event \mathcal{E}_i below.

The assumptions on the entries of A imply that

$$\mathbb{E} \|A_i\|_2^2 = n \quad \text{and} \quad \mathbb{E} \langle A_i, \mathbf{1} \rangle^2 = n.$$

Using Chebyshev's inequality, we see that

$$\mathbb{P} \{ \|A_i\|_2 > 2\delta^{-1}\sqrt{n} \} \leq \frac{\delta^2}{4} \quad \text{and} \quad \mathbb{P} \{ |\langle A_i, \mathbf{1} \rangle| > 2\delta^{-1}\sqrt{n} \} \leq \frac{\delta^2}{4}.$$

Then a union bound yields

$$\mathbb{P}(\mathcal{E}_i) \geq 1 - \frac{\delta^2}{2}.$$

By independence of the events \mathcal{E}_i , this implies

$$\mathbb{P}(\mathcal{E}) \geq \left(1 - \frac{\delta^2}{2}\right)^n \geq \exp(-\delta^2 n).$$

Next we apply Lemma 6.4, which states that the event

$$\mathcal{F} := \{ \|A\|_{\infty \rightarrow 2} \leq C\sqrt{n} \cdot \mathbb{E} \|A\|_{2 \rightarrow \infty} + C \mathbb{E} \|A\mathbf{1}\|_2 \}$$

is likely:

$$\mathbb{P}(\mathcal{F}) \geq 1 - \exp(-n).$$

It follows that

$$\mathbb{P}(\mathcal{E} \cap \mathcal{F}) \geq \exp(-\delta^2 n) - \exp(-n) \geq \frac{1}{2} \exp(-\delta^2 n).$$

It remains to note that by definition of \mathcal{E} and \mathcal{F} , the event $\mathcal{E} \cap \mathcal{F}$ implies the inequalities in (6.10). \square

6.4. Bounding $\infty \rightarrow 2$ norm with high probability. In the previous section, we were able to prove the optimal bounds

$$\|A\|_{2 \rightarrow \infty} \lesssim \sqrt{n} \quad \text{and} \quad \|A\|_{\infty \rightarrow 2} \lesssim n$$

for a random matrix A , but they only hold with exponentially small probability. We claim that the probability of success can be increased to almost 1 if we are allowed to remove a few columns of A . We already proved this fact for the $2 \rightarrow \infty$ norm in Lemma 5.1. It is time to handle the $\infty \rightarrow 2$ norm.

Theorem 6.6 (Bounding $\infty \rightarrow 2$ norm by removing a few columns). *Consider an $n \times n$ random matrix A with i.i.d. entries A_{ij} which have mean zero and unit variance and satisfy (5.1). Let $\varepsilon \in (0, 1/2]$. Then with probability at least $1 - 2\exp(-\varepsilon n)$, there exists a subset $J \in [n]$ with cardinality $|J| \leq \varepsilon n$ such that*

$$\|A_{J^c}\|_{\infty \rightarrow 2} \leq C\sqrt{\ln \varepsilon^{-1}} \cdot n.$$

Proof. Step 1: Defining the two key events. We will be interested in the two key events that suitably control the $2 \rightarrow \infty$ and $\infty \rightarrow 2$ norms of a random matrix. Thus, for a random matrix B and numbers $r, K \geq 0$, we define

$$\begin{aligned} \mathcal{E}_{2 \rightarrow \infty}(B, r, K) &:= \left\{ \exists J, |J| \leq r\varepsilon n : \|B_{J^c}\|_{2 \rightarrow \infty} \leq K\sqrt{\ln \varepsilon^{-1}} \cdot \sqrt{n} \right\}, \\ \mathcal{E}_{\infty \rightarrow 2}(B, r, K) &:= \left\{ \exists J, |J| \leq r\varepsilon n : \|B_{J^c}\|_{\infty \rightarrow 2} \leq K\sqrt{\ln \varepsilon^{-1}} \cdot n \right\}. \end{aligned}$$

In terms of these events, we want to show that

$$\mathbb{P}(\mathcal{E}_{\infty \rightarrow 2}(A, 1, C)^c) \leq 2 \exp(-\varepsilon n),$$

while Lemma 5.1 can be stated as

$$\mathbb{P}(\mathcal{E}_{2 \rightarrow \infty}(A, 1, C')) \geq 1 - \exp(-\varepsilon n).$$

for some absolute constant C' . Since the latter event is so likely, intersecting with it would not cause much harm. Indeed, we will show that the bad event

$$\mathcal{B} := \mathcal{E}_{2 \rightarrow \infty}(A, 1, C') \cap \mathcal{E}_{\infty \rightarrow 2}(A, 1, C)^c$$

satisfies

$$\mathbb{P}(\mathcal{B}) \leq \exp(-n/2). \quad (6.11)$$

This would finish the proof, since we would then have

$$\mathbb{P}(\mathcal{E}_{\infty \rightarrow 2}(A, 1, C)^c) \leq \exp(-n/2) + \exp(-\varepsilon n) \leq 2 \exp(-\varepsilon n)$$

as required.

Step 2: Symmetrization. As an intermediate step, let us bound the probability of a symmetrized version of \mathcal{B} , namely the event

$$\tilde{\mathcal{B}} := \mathcal{E}_{2 \rightarrow \infty}(\tilde{A}, 1, 2C') \cap \mathcal{E}_{\infty \rightarrow 2}(\tilde{A}, 1, C/2)^c$$

where

$$\tilde{A} := A - A'$$

and A' is an independent copy of the random matrix A . We claim that

$$\mathbb{P}(\tilde{\mathcal{B}}) \leq \exp(-n). \quad (6.12)$$

To prove this claim, choose a subset J , $|J| \leq \varepsilon n$, that minimizes $\|\tilde{A}_{J^c}\|_{2 \rightarrow \infty}$. Recall from (2.1) that the $2 \rightarrow \infty$ norm of a matrix is determined by the Euclidean norms of the columns and thus does not depend on the signs of the matrix elements. Thus J is independent of the signs of the elements of \tilde{A} . This makes it possible to use Lemma 6.2 for the matrix \tilde{A} and the random set J^c . It gives

$$\|\tilde{A}_{J^c}\|_{\infty \rightarrow 2} \lesssim \sqrt{n} \|\tilde{A}_{J^c}\|_{2 \rightarrow \infty} \quad (6.13)$$

with probability at least $1 - \exp(-n)$.

Then, turning to $\tilde{\mathcal{B}}$, we can bound its probability as follows:

$$\mathbb{P}(\tilde{\mathcal{B}}) \leq \mathbb{P}(\tilde{\mathcal{B}} \text{ and (6.13)}) + \exp(-n).$$

To prove the claim, it remains to check that $\tilde{\mathcal{B}}$ and (6.13) can not hold together. Assume they do; then

$$\|\tilde{A}_{J^c}\|_{\infty \rightarrow 2} \lesssim \sqrt{n} \cdot 2C' \sqrt{\ln \varepsilon^{-1}} \sqrt{n} \lesssim \sqrt{\ln \varepsilon^{-1}} \cdot n,$$

which contradicts the event $\mathcal{E}_{\infty \rightarrow 2}(\tilde{A}, 1, C/2)^c$ in the definition of $\tilde{\mathcal{B}}$ for a suitably chosen constant C . This completes the proof of the claim (6.12).

Step 3. Using the small-probability bounds. The last piece of information we will use is the conclusion of Lemma 6.5 for $\delta := 1/(2 \ln \varepsilon^{-1})$. It states that the good event

$$\mathcal{G} := \mathcal{E}_{2 \rightarrow \infty}(A', 0, C') \cap \mathcal{E}_{\infty \rightarrow 2}(A', 0, C/2)$$

is likely to happen:

$$\mathbb{P}(\mathcal{G}) \geq \frac{1}{2} \exp\left(-\frac{n}{4 \ln \varepsilon^{-1}}\right). \quad (6.14)$$

Note in passing that there is no guarantee that this statement would hold for the same constants C and C' as we chose in the definition of \mathcal{B} above. However, we can make this happen by adjusting these constants upwards as necessary. The reader can easily check both (6.12) and (6.14) would still hold after such an adjustment.

We claim that

$$\mathcal{B} \cap \mathcal{G} \subset \tilde{\mathcal{B}}. \quad (6.15)$$

To see this, recall that each of \mathcal{B} , \mathcal{G} and $\tilde{\mathcal{B}}$ is defined as an intersection of two events, one controlling $2 \rightarrow \infty$ norm and the other, $\infty \rightarrow 2$ norm. Thus it suffices to check the inclusion for each of these two parts separately. Namely, the claim (6.15) would follow at once if we show that

$$\begin{aligned} \mathcal{E}_{2 \rightarrow \infty}(A, 1, C') \cap \mathcal{E}_{2 \rightarrow \infty}(A', 0, C') &\subset \mathcal{E}_{2 \rightarrow \infty}(\tilde{A}, 1, 2C') \quad \text{and} \\ \mathcal{E}_{\infty \rightarrow 2}(A, 1, C)^c \cap \mathcal{E}_{\infty \rightarrow 2}(A', 0, C/2) &\subset \mathcal{E}_{\infty \rightarrow 2}(\tilde{A}, 1, C/2)^c. \end{aligned}$$

Both these inclusions are straightforward to check from the definitions of the events $\mathcal{E}_{2 \rightarrow \infty}$ and $\mathcal{E}_{\infty \rightarrow 2}$, remembering that $\tilde{A} = A - A'$ and using triangle inequality. This verifies the claim (6.15).

The event \mathcal{B} is determined by A , and \mathcal{G} is determined by A' only. Thus \mathcal{B} and \mathcal{G} are independent, and (6.15) gives

$$\mathbb{P}(\mathcal{B}) \mathbb{P}(\mathcal{G}) = \mathbb{P}(\mathcal{B} \cap \mathcal{G}) \leq \mathbb{P}(\tilde{\mathcal{B}}).$$

Thus, using (6.12) and (6.14), we conclude that

$$\mathbb{P}(\mathcal{B}) \leq \mathbb{P}(\tilde{\mathcal{B}})/\mathbb{P}(\mathcal{G}) \leq 2 \exp\left(-n + \frac{n}{4 \ln \varepsilon^{-1}}\right) \leq \exp(-n/2).$$

We have shown (6.11) and thus have completed the proof of the theorem. \square

7. FROM $\infty \rightarrow 2$ NORM TO THE OPERATOR NORM: CONTROLLING THE BOUNDED ENTRIES

In Theorem 6.6, we gave an optimal $O(n)$ bound for the $\infty \rightarrow 2$ norm of a random matrix with few removed columns. We will now convert this into an optimal $O(\sqrt{n})$ bound for the operator norm. This can be done by applying a form of Grothendieck-Pietsch theorem (see [11, Proposition 15.11]), which has been used recently in [10, section 3.2] in a similar context.

Theorem 7.1 (Grothendieck-Pietsch). *Let B be a $k \times m$ real matrix and $\delta > 0$. Then there exists $J \subset [m]$ with $|J| \leq \delta m$ such that*

$$\|B_{J^c}\| \leq \frac{2\|B\|_{\infty \rightarrow 2}}{\sqrt{\delta m}}.$$

Applying Theorem 6.6 followed by Grothendieck-Pietsch theorem, we obtain the following result.

Lemma 7.2 (Bounding the operator norm by removing a few columns). *Consider an $n \times n$ random matrix A with i.i.d. entries A_{ij} which have mean zero and unit variance and satisfy (5.1). Let $\varepsilon \in (0, 1]$. Then with probability at least $1 - 2\exp(-\varepsilon n/2)$, there exists a subset $J \in [n]$ with cardinality $|J| \leq \varepsilon n$ such that*

$$\|A_{J^c}\| \leq C \sqrt{\frac{\ln \varepsilon^{-1}}{\varepsilon}} \cdot \sqrt{n}.$$

Proof. Apply Theorem 6.6 for $\varepsilon/2$ instead of ε . We obtain a subset of columns $J_1 \subset [n]$, $|J_1| \leq \varepsilon n/2$, which satisfies

$$\|A_{J_1^c}\|_{\infty \rightarrow 2} \leq C \sqrt{\ln \varepsilon^{-1}} \cdot n \quad (7.1)$$

with probability at least $1 - 2\exp(-\varepsilon n/2)$.

Next apply Grothendieck-Pietsch Theorem 7.1 for the matrix $A_{J_1^c}$ and for $\delta = \varepsilon/2$. We obtain a further subset $J_2 \subset J_1^c$, $|J_2| \leq \delta |J_1^c| \leq \varepsilon n/2$, such that the removal of columns in both $J := J_1 \cup J_2$ leads to

$$\|A_{J^c}\| \leq \frac{2\|A_{J_1^c}\|_{\infty \rightarrow 2}}{\sqrt{\delta |J_1^c|}} \lesssim C \sqrt{\frac{\ln \varepsilon^{-1}}{\varepsilon}} \cdot \sqrt{n}.$$

In the last inequality, we used the bound (7.1) and that $\delta = \varepsilon/2$ and $|J_1^c| \geq n - \varepsilon n/2 \geq n/2$. The proof is complete. \square

We are ready to prove a partial case of Theorem 1.1, for the matrices whose entries are $O(\sqrt{n})$. It follows by applying Lemma 7.2 for A and A^\top separately, and then superposing the results.

Proposition 7.3. *Consider an $n \times n$ random matrix A with i.i.d. entries A_{ij} which have mean zero and unit variance and satisfy (5.1). Let $\varepsilon \in (0, 1]$. Then with probability at least $1 - 4\exp(-\varepsilon n/2)$, there exists an $\varepsilon n \times \varepsilon n$ submatrix of A such that replacing all of its entries with zero leads to a well-bounded matrix \tilde{A} :*

$$\|\tilde{A}\| \leq C \sqrt{\frac{\ln \varepsilon^{-1}}{\varepsilon}} \cdot \sqrt{n}.$$

Proof. Apply Lemma 7.2 for A and A^\top . We obtain that with probability at least $1 - 4\exp(-\varepsilon n/2)$, there exists sets I and J with at most εn indices in each, and such that

$$\|A_{[n] \times J^c}\| \lesssim \sqrt{\frac{\ln \varepsilon^{-1}}{\varepsilon}} \cdot \sqrt{n} \quad \text{and} \quad \|A_{I^c \times [n]}\| \lesssim \sqrt{\frac{\ln \varepsilon^{-1}}{\varepsilon}} \cdot \sqrt{n}. \quad (7.2)$$

We claim that $\tilde{A} := A_{(I \times J)^c}$ satisfies the conclusion of the proposition. The support of this matrix, $(I \times J)^c$, is a disjoint union of two sets, $[n] \times J^c$ and $I^c \times J$. Then, using the triangle inequality, we have

$$\|A_{(I \times J)^c}\| \leq \|A_{[n] \times J^c}\| + \|A_{I^c \times J}\|.$$

We already controlled the first term in (7.2). As for the second term, since adding columns can only increase the operator norm, we have $\|A_{I^c \times J}\| \leq \|A_{I^c \times [n]}\|$, which we also bounded in (7.2). The proof is complete. \square

8. CONTROLLING THE LARGE ENTRIES, AND COMPLETING THE PROOF OF THEOREM 1.1

In the previous section, we proved a partial case of Theorem 1.1 that controls relatively small entries of A , those of the order $O(\sqrt{n})$. Larger entries will be controlled in this section.

8.1. Bernoulli random matrices and random graphs. The following general lemma will help us analyze the patterns such large entries can form.

Lemma 8.1 (Bernoulli random matrix). *Let B be an $n \times n$ random matrix whose entries are independent Bernoulli random variables with mean p . Let $\varepsilon \in (0, 1/2]$. Consider the rows of B with more than $21pn + 2\ln \varepsilon^{-1}$ ones. Then with probability $1 - \exp(-\varepsilon n/2)$, these rows have at most εn ones altogether.*

To see the connection to our original problem, we will later choose the entries of B to be the indicators of the large entries of A .

Proof. Let S_i denote the number of ones in the i -th row of B . Then $\mathbb{E} S_i = pn$. A standard application of Chernoff's inequality shows that

$$\mathbb{P}\{S_i > t\} \leq e^{-2t} \quad \text{for } t \geq 21pn. \quad (8.1)$$

Let $K \geq 21pn$ be a number to be chosen later. (We will eventually choose K as $21pn + 2\ln \varepsilon^{-1}$ as in the statement of the lemma.) Define the random variables

$$X_i := S_i \mathbb{1}_{\{S_i > K\}}.$$

The quantity of interest is the total number of ones in the heavy rows, and it equals $\sum_{i=1}^n X_i$. To control this sum of independent random variables, we can use the standard Bernstein's trick (commonly called Chernoff's bound), where we use Markov's inequality after exponentiation. We obtain

$$\mathbb{P}\left\{\sum_{i=1}^n X_i > \varepsilon n\right\} \leq e^{-\varepsilon n} \mathbb{E} \exp\left(\sum_{i=1}^n X_i\right) = \left[e^{-\varepsilon} \mathbb{E} e^{X_1}\right]^n, \quad (8.2)$$

where the last equality follows by independence and identical distribution. Now, by definition of X_1 we have

$$\begin{aligned}
\mathbb{E} e^{X_1} &= \mathbb{E} e^{X_1} \mathbf{1}_{\{X_1=0\}} + \mathbb{E} e^{X_1} \mathbf{1}_{\{X_1 \neq 0\}} \leq 1 + \mathbb{E} e^{S_1} \mathbf{1}_{\{S_1 > K\}} \\
&= 1 + \int_{e^K}^{\infty} \mathbb{P} \{e^{S_1} > u\} du \\
&= 1 + \int_K^{\infty} \mathbb{P} \{S_1 > t\} e^t dt \quad (\text{by a change of variables}) \\
&\leq 1 + \int_K^{\infty} e^{-2t} e^t dt \quad (\text{using (8.1) for } t \geq K \geq 21pn) \\
&= 1 + e^{-K} \leq \exp(e^{-K}).
\end{aligned}$$

Substituting this bound into (8.2), we conclude that

$$\mathbb{P} \left\{ \sum_{i=1}^n X_i > \varepsilon n \right\} \leq \exp [(-\varepsilon + e^{-K})n] \leq \exp(-\varepsilon n/2),$$

if we choose K so that $e^{-K} \leq \varepsilon/2$. To finish the proof, recall that our argument works if K satisfies the two conditions: $K \geq 21pn$ and $e^{-K} \leq \varepsilon/2$. We thus choose $K := 21pn + 2 \ln \varepsilon^{-1}$ and complete the proof. \square

Corollary 8.2 (Bernoulli random matrix). *Let B be an $n \times n$ random matrix whose entries are independent Bernoulli random variables with mean p . Let $\varepsilon \in (0, 1]$. Then with probability at least $1 - 2 \exp(-\varepsilon n/4)$, there exists an $\varepsilon n \times \varepsilon n$ submatrix of B such that replacing all of its entries with zero leads to a matrix \tilde{B} whose rows and columns have at most $21pn + 4 \ln \varepsilon^{-1}$ ones each.*

Proof. Apply Lemma 8.1 for B and B^\top with $\varepsilon/2$ instead of ε , and take the intersection of the two good events. With the required probability, we obtain a set of εn bad entries of B whose removal makes all rows and columns of B contain at most $21pn + 2 \ln \varepsilon^{-1}$ ones. It remains to note that these εn entries can be trivially placed in some $\varepsilon n \times \varepsilon n$ submatrix of B . \square

Remark 8.3 (Random graphs). It is not difficult to obtain a version of Corollary 8.2 for symmetric random matrices. This version can be interpreted as a statement about Erdős-Rényi random graphs $G(n, p)$, with B playing the role of the adjacency matrix. It states that with high probability, one can make all degrees of a $G(n, p)$ random graph bounded by $O(pn + \ln \varepsilon^{-1})$ after removing the internal edges from a sub-graph with εn vertices.

8.2. Moderately large entries. We will use Corollary 8.2 to deduce Theorem 1.1 for matrices with moderately large entries. Namely, we assume here that all entries of A satisfy

$$A_{ij} = 0 \quad \text{or} \quad \frac{\sqrt{n}}{2} \leq |A_{ij}| \leq \frac{5\sqrt{n}}{\sqrt{\varepsilon}}. \quad (8.3)$$

Proposition 8.4. *Consider an $n \times n$ random matrix A with i.i.d. entries which satisfy $\mathbb{E} A_{ij}^2 \leq 1$ and (8.3). Let $\varepsilon \in (0, 1/2]$. Then with probability at least $1 - 2\exp(-\varepsilon n/4)$, there exists an $\varepsilon n \times \varepsilon n$ submatrix of A such that replacing all of its entries with zero leads to a well-bounded matrix \tilde{A} :*

$$\|\tilde{A}\| \leq \frac{C \ln \varepsilon^{-1}}{\sqrt{\varepsilon}} \cdot \sqrt{n}. \quad (8.4)$$

Proof. Consider the matrix B whose elements are indicators of moderately large entries of A , i.e.

$$B_{ij} := \mathbf{1}_{\{A_{ij} \neq 0\}}.$$

Then B_{ij} are i.i.d. Bernoulli random variables with mean

$$p := \mathbb{E} B_{ij} = \mathbb{P}\{A_{ij} \neq 0\} \leq \mathbb{P}\left\{|A_{ij}| \geq \frac{\sqrt{n}}{2}\right\} \leq \frac{2}{n}. \quad (8.5)$$

(In the last inequality, we used Chebyshev's inequality and the assumption $\mathbb{E} A_{ij}^2 \leq 1$.) Corollary 8.2 applied to B gives us an $\varepsilon n \times \varepsilon n$ submatrix of A such that the number of non-zero elements in every row and column of \tilde{A} (obtained by zeroing out the elements of A outside that submatrix) is bounded by

$$21pn + 4 \ln \varepsilon^{-1} \lesssim \ln \varepsilon^{-1}, \quad (8.6)$$

where we used (8.5) in the last bound.

Moreover, assumption (8.3) shows that all entries of \tilde{A} are bounded in absolute value by $5\sqrt{n}/\sqrt{\varepsilon}$. This and (8.6) imply that the ℓ_1 norm of all rows \tilde{A}_i and columns \tilde{A}^j can be bounded as follows:

$$\max_{i,j} \left(\|\tilde{A}_i\|_1, \|\tilde{A}^j\|_1 \right) \lesssim \frac{\sqrt{n}}{\sqrt{\varepsilon}} \cdot \ln \varepsilon^{-1}.$$

Applying Lemma 3.1 leads to (8.4). \square

8.3. Very large entries. Finally, we will need to prove Theorem 1.1 for very large entries – now we assume that all entries of A satisfy

$$A_{ij} = 0 \quad \text{or} \quad |A_{ij}| > \frac{5\sqrt{n}}{\sqrt{\varepsilon}}. \quad (8.7)$$

There are typically very few such entries, as the following simple result shows.

Lemma 8.5 (Few very large entries). *Consider an $n \times n$ random matrix A with i.i.d. entries which satisfy $\mathbb{E} A_{ij}^2 \leq 1$ and (8.3). Let $\varepsilon \in (0, 1/2]$. Then with probability at least $1 - \exp(-\varepsilon n)$, the matrix A has at most εn non-zero entries.*

Proof. Using Chebyshev's inequality and the assumption that $\mathbb{E} A_{ij}^2 \leq 1$, we see that the probability that a given entry is nonzero is

$$\mathbb{P}\{A_{ij} \neq 0\} \leq \mathbb{P}\left\{|A_{ij}| > \frac{5\sqrt{n}}{\sqrt{\varepsilon}}\right\} \leq \frac{\varepsilon}{25n}.$$

Thus the expected number of non-zero entries in A is at most $\varepsilon n/25$. A standard application of Chernoff's inequality gives

$$\mathbb{P}\{A \text{ has more than } \varepsilon n \text{ nonzero entries}\} \leq e^{-\varepsilon n}.$$

The proof is complete. \square

Since a set of εn indices can be always placed in an $\varepsilon n \times \varepsilon n$ submatrix, we can state Lemma 8.5 as follows.

Corollary 8.6 (Few very large entries). *Consider an $n \times n$ random matrix A with i.i.d. entries which satisfy $\mathbb{E} A_{ij}^2 \leq 1$ and (8.3). Let $\varepsilon \in (0, 1/2]$. Then with probability at least $1 - \exp(-\varepsilon n)$, all non-zero entries of A are contained in an $\varepsilon n \times \varepsilon n$ submatrix.*

8.4. Proof of Theorem 1.1. The proof follows simply by combining Proposition 7.3 for small entries of A , Proposition 8.4 for moderately large entries, and Corollary 8.6 for very large entries. Technically, we decompose A into a sum of three matrices

$$A = A_1 + A_2 + A_3$$

which contain small, moderately large and very large entries of A respectively. Then we apply the results quoted above with $\varepsilon/3$ instead of ε , and take the intersection of the three good events. At the end, we embed the three $\varepsilon n/3 \times \varepsilon n/3$ resulting submatrices into one $\varepsilon n \times \varepsilon n$ submatrix. Theorem 1.1 follows.

9. GLOBAL PROBLEM: PROOF OF THEOREM 1.3

In this section we prove Theorem 1.3, which states that either nonzero mean or infinite second moment make it impossible to repair the matrix norm by removing a small submatrix. We will first prove a non-asymptotic version of this result. Once this is done, an application of Borel-Cantelli Lemma will quickly yield Theorem 1.3.

Proposition 9.1 (Global problem: non-asymptotic regime). *Consider an $n \times n$ random matrix A whose entries are i.i.d. random variables that have either nonzero mean or infinite second moment, and let $\varepsilon \in (0, 1)$. Then, for any $M > 0$ there exists n_0 that may depend only on ε , M and the distribution of the entries, and such that for any $n > n_0$ the following event holds with probability at least $1 - e^{-n}$: every $(1 - \varepsilon)n \times (1 - \varepsilon)n$ submatrix A' of A satisfies*

$$\|A'\| \geq M\sqrt{n}.$$

Before we prove this proposition, let us pause to see its connection to the matrix \tilde{A}_n of Theorem 1.3. Proposition 9.1 yields that this matrix satisfies

$$\|\tilde{A}_n\| \geq M\sqrt{n}.$$

Indeed, modifying an $\varepsilon n \times \varepsilon n$ submatrix always leaves some $(1 - \varepsilon)n \times (1 - \varepsilon)n$ submatrix A' intact, so we can apply Proposition 9.1 for that submatrix.

9.1. Infinite second moment. Here we will prove the part of Proposition 9.1 about infinite second moment; the case of nonzero mean will be treated in Section 9.2. Let us start with the following lemma which will help us treat a fixed submatrix.

Lemma 9.2. *Consider an $m \times m$ random matrix B whose entries are i.i.d. random variables with infinite second moment. Then, for any $M > 0$ there exists m_0 that may depend only on M and the distribution of the entries, and such that for any $m > m_0$ we have*

$$\|B\| \geq M\sqrt{m}$$

with probability at least $1 - \exp(-M^2 m)$.

Proof. By assumption, we have $\mathbb{E} B_{ij}^2 = \infty$. Therefore, for any $M > 0$ one can find a truncation level K that depends only on M and the distribution, and such that the truncated random variables

$$\bar{B}_{ij} := B_{ij} \mathbf{1}_{|B_{ij}| \leq K} \quad \text{satisfy} \quad \mathbb{E} \bar{B}_{ij}^2 \geq 2M^2. \quad (9.1)$$

(This follows easily from Lebesgue's monotone convergence theorem.)

Consider the matrix \bar{B} with entries \bar{B}_{ij} . We have

$$\|B\| \geq \frac{1}{\sqrt{m}} \|B\|_F \geq \frac{1}{\sqrt{m}} \|\bar{B}\|_F.$$

Then we bound the failure probability as follows:

$$\begin{aligned} \mathbb{P} \{ \|B\| < M\sqrt{m} \} &\leq \mathbb{P} \{ \|\bar{B}\|_F < Mm \} = \mathbb{P} \left\{ \sum_{i,j=1}^m \bar{B}_{ij}^2 < M^2 m^2 \right\} \\ &\leq \mathbb{P} \left\{ \sum_{i,j=1}^m (\bar{B}_{ij}^2 - \mathbb{E} \bar{B}_{ij}^2) < -M^2 m^2 \right\} \end{aligned}$$

where we used (9.1) in the last step.

Apply Hoeffding's inequality for the random variables \bar{B}_{ij}^2 and use that they are bounded by K^2 by construction. The probability above gets bounded by

$$\exp \left(- \frac{M^4 m^2}{2K^2} \right).$$

If $m > 2K^2/M^2 = m_0$, this probability can be further bounded by $\exp(-M^2 m)$, as claimed. \square

Proof of Proposition 9.1 for infinite second moment. We can assume without loss of generality that M is large enough depending on ε . (Indeed, once the conclusion of the proposition holds for one value of M it automatically holds for all smaller values.)

Apply Lemma 9.2 for an $m \times m$ matrix A'_n with $m = (1 - \varepsilon)n$, and then take a union bound over all $\binom{n}{m}^2$ possible choices of such submatrices. It follows that the conclusion of Proposition 9.1 holds with probability at least

$$1 - \binom{n}{m}^2 \exp(-M^2 m).$$

By Stirling's approximation, we have $\binom{n}{m} \leq (en/m)^m$. Using this and substituting $m = (1 - \varepsilon)n$, we bound the probability below by

$$1 - \exp \left[\left(2 \log \frac{e}{1 - \varepsilon} - M^2 \right) (1 - \varepsilon)n \right].$$

If the value of M is sufficiently large depending on ε , this probability is larger than $1 - \exp(-n)$, as claimed. Proposition 9.1 for infinite second moment is proved. \square

9.2. Nonzero mean. Now we will prove the part of Proposition 9.1 about nonzero mean. We can assume here that the second moment of the entries A_{ij} is finite, as the opposite case was treated in Section 9.1. As before, we will first focus on one submatrix. In the following lemma we make an extra boundedness assumption, which we will get rid of using truncation later.

Lemma 9.3. *Consider an $m \times m$ random matrix B whose entries are i.i.d. random variables that satisfy*

$$\mathbb{E} B_{ij} = \mu > 0, \quad \mathbb{E} B_{ij}^2 \leq \sigma^2, \quad |B_{ij}| \leq K\sqrt{m} \text{ a.s.}$$

Then, for any $M > 0$ there exists m_0 that may depend only on μ , σ , K and M , and such that for any $m > m_0$ we have

$$\|B\| \geq \frac{\mu m}{2}$$

with probability at least $1 - \exp(-M^2 m)$.

Proof. Notice that

$$\|B\| \geq \frac{1}{m} \sum_{i,j=1}^m B_{ij}.$$

(To check this inequality, recall that $\|B\| \geq x^\top B x$ for any unit vector x ; use this for the vector x whose all coordinates equal $1/\sqrt{m}$.) Then we can bound the failure probability as follows:

$$\mathbb{P} \left\{ \|B\| < \frac{\mu m}{2} \right\} \leq \mathbb{P} \left\{ \sum_{i,j=1}^m B_{ij} < \frac{\mu m^2}{2} \right\} \leq \mathbb{P} \left\{ \sum_{i,j=1}^m (B_{ij} - \mathbb{E} B_{ij}) < -\frac{\mu m^2}{2} \right\}$$

where we used that $\mathbb{E} B_{ij} = \mu$ in the last step.

Apply Bernstein's inequality for the random variables B_{ij} and use that they have variance at most σ^2 and are bounded by $K\sqrt{m}$ by assumption.

The failure probability gets bounded by

$$\exp\left(-\frac{\mu^2 m^4/8}{\sigma^2 m^2 + K\sqrt{m}/3}\right).$$

If m is large enough depending μ , σ , K and M , then this probability can be further bounded by $\exp(-M^2 m)$, as claimed. \square

Next, we will use truncation to get rid of the boundedness assumption in Lemma 9.3 and thus prove the following.

Lemma 9.4. *Consider an $m \times m$ random matrix B whose entries are i.i.d. random variables that satisfy*

$$\mathbb{E} B_{ij} = \mu > 0, \quad \mathbb{E} B_{ij}^2 \leq \sigma^2.$$

Then, for any $M > 0$ there exists m_0 that may depend only on μ , σ , K , M and the distribution of the entries, and such that for any $m > m_0$ we have

$$\|B\| \geq M\sqrt{m} \tag{9.2}$$

with probability at least $1 - \exp(-M^2 m)$.

Proof. Choosing m_0 large enough depending on M and the distribution of B_{ij} , we can make sure that for any $m \geq m_0$ the truncated random variables

$$\bar{B}_{ij} := B_{ij} \mathbb{1}_{|B_{ij}| \leq M\sqrt{m}} \quad \text{satisfy} \quad \mathbb{E} \bar{B}_{ij} \geq \mathbb{E} B_{ij} - \frac{\mu}{2} = \frac{\mu}{2}.$$

(This follows easily from Lebesgue's monotone convergence theorem.)

Let us consider the event that all entries of B are appropriately bounded:

$$\mathcal{E} := \{|B_{ij}| \leq M\sqrt{m} \text{ for all } i, j \in [n]\}.$$

Suppose for a moment that (9.2) fails, so we have $\|B\| < M\sqrt{m}$. Since the inequality $\|B\| \geq \max_{i,j} |B_{ij}|$ is always true, the event \mathcal{E} must hold in this case. This in turn implies that the truncation has no effect on the entries, i.e. $\bar{B}_{ij} = B_{ij}$ for all i, j .

We have shown that in the event of the failure of (9.2), we may automatically assume that the entries of B are appropriately bounded. Therefore the failure probability satisfies

$$\mathbb{P}\{\|B\| < M\sqrt{m}\} = \mathbb{P}\{\|\bar{B}\| < M\sqrt{m}\}$$

where \bar{B} denotes the matrix with the truncated entries \bar{B}_{ij} . It remains to apply Lemma 9.3 for the random matrix \bar{B} , noting that truncation may only decrease the second moment. The failure probability gets bounded by $\exp(-M^2 m)$, as claimed. \square

Proof of Proposition 9.1 for non-zero mean. As we mentioned in the beginning of this section, we can assume that the entries B_{ij} have finite second moment σ^2 . Then the conclusion of the proposition follows by exact same union bound argument as in the end of Section 9.1 (just use Lemma 9.4 instead of Lemma 9.2 there.) \square

9.3. Proof of Theorem 1.3. We will prove a stronger fact that

$$\min \frac{\|A'_n\|}{\sqrt{n}} \rightarrow \infty \quad \text{as } n \rightarrow \infty \quad \text{almost surely,} \quad (9.3)$$

where the minimum is taken over all $(1-\varepsilon)n \times (1-\varepsilon)n$ submatrices A'_n of A_n . As we mentioned below Proposition 9.1, this would imply the conclusion of Theorem 1.3, since modifying an $\varepsilon n \times \varepsilon n$ submatrix leaves some $(1-\varepsilon)n \times (1-\varepsilon)n$ sub-matrix intact.

Fix any $M > 0$ and consider the events

$$\mathcal{E}_n := \left\{ \min \frac{\|A'_n\|}{\sqrt{n}} \geq M \right\}, \quad n = 1, 2, \dots$$

where the minimum has the same meaning as before. By Proposition 9.1, there exists n_0 such that

$$\mathbb{P}(\mathcal{E}_n^c) \leq e^{-n} \text{ for all } n > n_0.$$

In particular, the series $\sum_{n=1}^{\infty} \mathbb{P}(\mathcal{E}_n^c)$ converges. Borel-Cantelli lemma then implies that the probability that infinitely many \mathcal{E}_n^c occur is 0. Equivalently, with probability 1 there exists N such that \mathcal{E}_n hold for all $n \geq N$.

We have shown that for any $M > 0$, with probability 1 there exists N such that

$$\min \frac{\|A'_n\|}{\sqrt{n}} \geq M \quad \text{for all } n \geq N.$$

Intersecting these almost sure events for $M = 1, 2, \dots$, we conclude (9.3). Theorem 1.3 is proved. \square

10. FURTHER QUESTIONS

Several extensions of Theorem 1.1 seem plausible.

1. It is natural to expect a version Theorem 1.1 even if the entries of A are *not identically distributed*. Our argument relies on the identical distribution in several places, including discretization arguments (proof of Theorem 4.2) and symmetrization (proofs of Lemmas 6.3 and 6.4).
2. A version of Theorem 1.1 should hold for *symmetric matrices* A with independent entries on and above the diagonal. A simplest way to get this result would be to use Theorem 1.1 to control the parts of A above and below the diagonal separately, and then combine them. However, for this argument one would need a version of Theorem 1.1 for non-identical distributed entries.
3. Unlike Feige-Ofek's result [6] mentioned in Section 1.4, Theorem 1.1 does not indicate what sub-matrix should be removed to improve the norm; it is rather an existential result. It would be nice to have an *explicit description of a submatrix to be removed*.
4. It would be good to *remove the logarithmic factor* $\ln \varepsilon^{-1}$ from the bound in Theorem 1.1, or to show that this factor is necessary. Such bound would be optimal up to an absolute constant factor.

5. Finally, while Remark 1.2 states that *the dependence on ε* in Theorem 1.1 is optimal in general, this dependence might be dramatically improved under a natural boundedness assumption. Namely, suppose that the entries of A are $O(\sqrt{n})$ almost surely. (In fact, most of the proof – until Section 8 – was done under this additional assumption.) In this case, is the dependence of the norm on ε *logarithmic* in Theorem 1.1, i.e.

$$\|\tilde{A}\| \leq C \ln(\varepsilon^{-1}) \sqrt{n}? \quad (10.1)$$

In fact, for the partial case of Bernoulli matrices such that $np = c_0 = \text{const}$ (where p is a probability of a non-zero entry) this bound can be quickly deduced from Corollary 8.2.

Indeed, after renormalization that imposes matrix elements to have variance one (so we deal with the scaled Bernoulli matrix with $B_{ij} = O(p^{-1/2})$), we can see that such matrices satisfy the boundedness assumption, as $B_{ij} = O(p^{-1/2}) = O(\sqrt{n}/\sqrt{c_0}) = O(\sqrt{n})$. Then, by Corollary 8.2 after a deletion of $\varepsilon n \times \varepsilon n$ submatrix we get a matrix \tilde{B} with all rows and columns having at most

$$21pn + 4 \ln \varepsilon^{-1} \leq 100c_0 \ln \varepsilon^{-1}$$

non-zero elements of order $O(\sqrt{n})$. Hence,

$$\max_{i,j} \left(\|\tilde{B}_i\|_1, \|\tilde{B}^j\|_1 \right) \lesssim \sqrt{n} \cdot \ln \varepsilon^{-1}.$$

Applying Lemma 3.1 leads to (10.1).

REFERENCES

- [1] N. Alon, A. Naor, *Approximating the cut-norm via Grothendiecks inequality*. SIAM J. Comput. 35, no. 4, 787–803, 2006.
- [2] Z. Bai, Y. Yin, *Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a Wigner matrix*. Ann. Probab. 16, no. 4, 1729–1741, 1988.
- [3] B. Bollobas, S. Janson, O. Riordan, *The cut metric, random graphs, and branching processes*. J. Statist. Phys. 140:2, 289–335, 2010.
- [4] Z. Bai, J. Silverstein, Y. Yin, *A note on the largest eigenvalue of a large-dimensional sample covariance matrix*. J. Multivariate Anal., 26, 166–168, 1988.
- [5] R. van Handel, *On the spectral norm of Gaussian random matrices*. Trans. Amer. Math. Soc., to appear.
- [6] U. Feige, E. Ofek, *Spectral techniques applied to sparse random graphs*. Random Structures Algorithms 27(2), 251–275, 2005.
- [7] M. Krivelevich, B. Sudakov, *The largest eigenvalue of sparse random graphs*. Combin. Probab. Comput., 12:61–72, 2003.
- [8] R. Latała, *Some estimates of norms of random matrices*. Proc. Amer. Math. Soc. 133 (2005), 1273–1282.
- [9] A. E. Litvak, S. Spector, *Quantitative version of a Silversteins result*. GAFA, Lecture Notes in Math., 2116 (2014), 335–340.
- [10] C. Le, E. Levina, R. Vershynin, *Concentration and regularization of random graphs*. Random Structures and Algorithms, to appear.
- [11] M. Ledoux, M. Talagrand, *Probability in Banach spaces: Isoperimetry and processes*. Springer-Verlag, Berlin, 1991.

- [12] S. Mendelson, G. Paouris, *On the singular values of random matrices*. Journal of EMS, 16 (2014), no 4, 823–834.
- [13] E. Rebrova, K. Tikhomirov, *Refined ε -nets and invertibility of random square matrices with i.i.d. heavy-tailed entries*. Preprint, arXiv:1508.06690
- [14] A. Pietsch, *Operator Ideals*. North-Holland Amsterdam, 1978.
- [15] G. Pisier, *Factorization of linear operators and geometry of Banach spaces*. Number 60 in CBMS Regional Conference Series in Mathematics. AMS, Providence, 1986.
- [16] Y. Seginer, *The expected norm of random matrices*. Combin. Probab. Comput. 9 (2000), no. 2, 149–166.
- [17] J. A. Tropp, *An Introduction to Matrix Concentration Inequalities*. Found. Trends Mach. Learning, Vol. 8, num. 1-2, pp. 1–230, 2015.
- [18] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*. Compressed sensing, 210–268, Cambridge Univ. Press, Cambridge, 2012.
- [19] E. Wolfstetter, *Topics in Microeconomics*. 1st ed. Cambridge: Cambridge University Press, 1999.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF MICHIGAN, 530 CHURCH ST, ANN ARBOR, MI 48109, U.S.A.

E-mail address: erebrova@umich.edu, romanv@umich.edu